이미지 속 문자열 탐지에 대한 YOLO와 EAST 신경망의 성능 비교

박찬용*, 이규현*, 임영민**, 정승대**, 조영혁**, 김진욱*
*경북대학교 IT대학 컴퓨터학부, **주식회사 투아트
bw_yong13@tuat.kr, prodzpod@protonmail.com, youngmin@tuat.kr, sdjeong@tuat.kr,
Philip@tuat.kr, deepkaki@knu.ac.kr

A Comparison of Deep Neural Network based Scene Text Detection with YOLO and EAST

Chan-Yong Park*, Gyu-Hyun Lee*, Young-Min Lim**, Seung-Dae Jeong**, Young-Heuk Cho**, Jin-Wook Kim*
*School of Computer Science and Engineering, KyungPook National University, **TUAT Corp.

요 약

본 논문에서는 최근 다양한 분야에서 많이 활용되고 있는 YOLO와 EAST 신경망을 이미지 속 문자열 탐지문제에 적용해보고 이들의 성능을 비교분석 해 보았다. YOLO 신경망은 v3 이전 모델까지는 이미지 속 문자영역 탐지에 낮은 성능을 보인다고 알려졌으나, 최근 출시된 YOLOv4와 YOLOv5의 경우 다양한 형태의 이미지 속에 있는 한글과 영문 문자열 탐지에 뛰어난 성능을 보여줌을 확인하고 향후 문자 인식 분야에서 많이 활용될 것으로 기대된다.

1. 서론

복잡한 배경속의 문자열을 탐지하는 기술은 자율주행 자동차나 로봇에 적용하는 시각 기술의 일환으로 발전되어 왔다. 이미지 속 문자탐지는 실시간 번역, 이미지 검출, 영상 파싱, 위치정보추출 그리고 시각장애인 길안내 등 수많은 응용분야 때문에 컴퓨터 비전 연구에서 크게 주목을 받고 있다[1]. 하지만 이미지 속 문자 인식은 인식할 문자열이 이미지 속 어디 있는지 탐지가 어렵고, 카메라 성능과 초점에 따라 이미지 퀄러티가 영향을 많이 받으며, 텍스트가 평행한 각도가 아니고 심지어 이미지가 회전 될 수도 있으며 다양한 폰트들이 존재하기에 기존 OCR과는 차원이 다른 난이도를 가진다[2].

일반적으로 이미지 속 문자인식은 문자열 탐지와 문자인식 이렇게 2단계로 구분되는데, 문자인식만큼 문자열 탐지도 어려운 미션이기에 다양한 방법들이 시도되어 왔다. 전통적인 문자열 탐지방법들은 문자/단어 후보 생성, 후보 필터링 및 그룹핑 등과 같은 다수의 처리단계를 둔다. 하지만, 이런 접근법들은 각 단계별로 수많은 파라미터들을 튜닝하고 휴리스틱 룰들을 적용하면서도 결국에는 전반적인 성능저하를 보여준다고 평가 받는다[3].

최근 SSD, Faster R-CNN그리고 FCN과 같은 객체 탐지/분할 방법을 채택한 딥러닝 기반 문자열 탐지방법들이 제시되고 있다[4,5]. 이들 방법은 이미지 속에서 단어단위 경계박스(word level bounding box)를 찾기 위해 신경망을 학습하여 주목할 만한 결과들을 보여주고 있다.

본 논문에서는 현재 시각 장애인들의 시각보조 앱으로 서비스 중인 설리번플러스에 적용할 목적으로 YOLO(You Only Look Once)와 EAST(An Efficient and Accurate Scene Text Detector) 신경망을 이용하여 다양한 이미지 속 문자열을 탐지하고 이들 신경망의 문자 탐지 성능에 대한 데이터를 제시한다.

최신 YOLO 신경망을 문자 탐지에 특화하여 적용한 사례가 드물며 특히 한글 문자 탐지에 대해서 집중적으로 테스트한 사례는 전무하기에 본 논문의 실험 결과는 향후 유사한 문제에 YOLO와 EAST를 적용하고자 할 때 참고할 수 있을 것이다.

다음 장에서는 시각장애인의 시각보조 앱에 대해 설명하고 이어서 YOLO와 EAST 신경망의 주요 특징 그리고 이들 각각의 신경망 모델에 실제 데이터를 적용한 실험 결과에 대해 설명하고 결론을 도출한다.

2. 시각장애인 시각보조 앱

설리번플러스 서비스는 시각 장애인들을 위해 스마트폰 카메라를 이용하여 이미지를 인식한 후 음성으로 내용을 전달해주는 시각장애인용 시각보조 앱이다. 설리번플러스의 주요 기능으로는 이미지 캡셔닝 기술을 이용한 이미지 인식, 이미지 속 문자를 인식하여 읽어주는 문자인식 기술 그리고 사람의 얼굴 인식 기술이 있다. 실제 사용자들의 이용 패턴을 보면, 문자 인식에 대한 사용 빈도가 가장 높은데 그만큼 사람들이 문자를 통해 취득하는 정보가 많기 때

문이다.

시각장애인들은 스마트폰 카메라를 통해 대상을 인식하고자 할 때 카메라 화각 안에 인식할 대상을 정확하게 담는 것이 가장 큰 이슈가 된다. 정안인들은 어려움 없이 카메라 화각에 촬영하고자 하는 대상을 담을 수 있지만, 앞이 보이지 않는 시각 장애인들은 카메라 화각에 대상을 담는 부분이 가장 어려운 문제이다. 이런 문제 때문에 대부분의 시각 장애인 시각 보조 앱들의 경우, 카메라 화각 안에 대상을 담을 수 있도록 소리와 진동으로 사용자에게 피드백을 해주는 기능을 가지고 있다. 이 때문에 설리번 플러스의 문자 인식 기능의 경우도, 이미지 속 문자를 인식하는 부분과 문자열을 탐지하는 부분 둘 다 중요한 성능 지표로 두고 성능 향상을 위해 노력해오고 있다.

특히 시각 장애인들이 스마트폰 카메라를 이용하여 문자를 탐지하는 상황은 사용자에게 빠른 피드백을 제공하여야 해서 스마트폰 환경에서 동작하는 엔진의 성능이 무엇보다 중요하다. 서비스 사용의 특성상 문자열 탐지 기능은 모바일 단말기 단에서 빠르고 가볍게 동작하여야 한다. 이에 가볍고 성능이 뛰어난 문자 탐지 엔진 개발을 위해 본 연구를 진행하게 되었다.

3. YOLO와 EAST 신경망을 활용한 문자열 탐지

EAST는 FCN(Fully Convolutional Network)을 이용하는 빠르고 정확한 2단계의 문자열 탐지 파이프라인으로 단어 단위 혹은 문자 단위로 문자열을 추출한다[6]. 반면에 YOLO는 이미지를 S X S개의 그리드로 분할하고 각각의 그리드 내 객체의 신뢰도를 계산한 뒤 그리드를 합치는 과정에서 경계상자의 위치를 조정하면서 객체 추출 정확도를 높이는데 매우 빠르고 강력한 성능을 자랑한다[7]. YOLO의 경우 v3모델까지는 문자열 탐지에 좋은 평가를 보여주지 못했으나[8] 최근 출시된 YOLOv4와 v5에서는 이전 모델보다 월등히 개선된 특성들을 보여주고 있다. 이에 본 연구에서는 YOLOv4와 v5가 문자열 탐지에도 충분히 효과적으로 사용될 수 있을 것으로 기대하면서 문자열 탐지에 적용해 보았다. 그리고 그 결과를 EAST 신경망 모델과 비교하여 제시한다. 문자열 탐지, 특히 한글에 대해 이 두 모델의 성능비교에 대한 자료가 전무한 상황이기에 본 연구는 의의를 가진다고 본다.

주어진 이미지 속 문자열에 신경망으로 테두리를 형성(Boxing)하는 방법으로 AABB(Axis-Aligned Bounding Box)와 RBOX(Rotated Bounding Box) 구조가 있는데, AABB는 임의 형태 사각 내의 각 포인트에서 직사각형 4변까지의 거리 정보로 정렬된 사각형 구조 이고 RBOX 는 이러한 AABB 정보에 사각형의 기울어진 각도 정보도 포함하여 기울어진 문자 영역을 좀 더 정확하게 탐지하는 방법이다. 본 논문에서는 EAST 모델은 RBOX 구조를 사용하고, YOLO 모델은 AABB 구조를 사용한다.

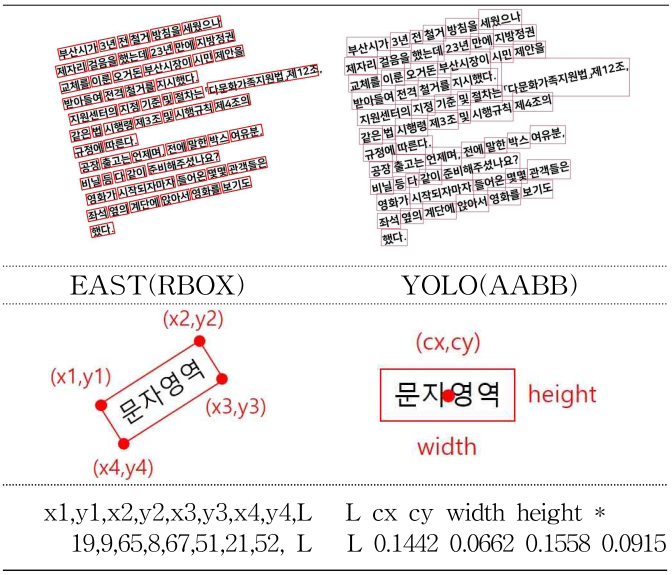


그림 1. RBOX 와 AABB Boxing 예시

* YOLO의 AABB구조는 전체 이미지의 가로, 세로 길이를 1로 봤을 때 상대적인 값으로 위치와 너비, 높이를 표시함

4. 실험 및 결과

본 실험에 사용한 신경망 모델별 이미지 크기는 EAST 512×512, YOLOv4 608x608 그리고 YOLOv5 640×640 픽셀 사이즈(신경망 입력 크기) 이며, RGB 32비트 이미지이다. 또한 실험에 사용한 장비의 경우 OS는 Ubuntu 18.04 LTS, GPU는 GeForce RTX 2080 TI 11GB이다. 설리번플러스 서비스는 스마트폰 카메라로 촬영한 영상속의 문자열을 탐지하는 기술이기에, 학습 데이터를 생성 시 이미지의 다양한 변형에 대한 특성을 감안하여 학습 데이터를 생성 하였으며 특히 한국지능정보사회진흥원의 AI Hub에서 제공하는 데이터 셋을 바탕으로 문자영역 탐지 실험을 진행하였다.

AI Hub는 국내 중소벤처기업, 연구소, 개인 등이 높은 비용과 투입시간으로 인해 자체적으로 확보하기 어려운 양질의 대용량 인공지능 학습용 데이터를 제공해 주는 곳으로 AI 데이터 종류로는 공공/법률, 과학기술/정보통신, 교육/문화/스포츠, 교통/물류, 농업/축산/수산/임업/식품, 보건/복지/의료, 재난/안전, 환경/기후 등이 있다.

본 논문에서는 두 가지 데이터 세트를 활용하였는데 첫 번째는 문서에 있는 문자를 감지하는 형태이고 두 번째는 실생활에서 볼 수 있는 간판, 책 표지 등 과 같이 다양한 배경과 조합된 문자를 감지하는 형태이다. 첫 번째 문서에 있는 문자 데이터 세트는 한국어-영어 번역 말뭉치의 문어체 뉴스 데이

터 세트(영문 - 10,000 문장, 한글 - 800,000 문장)를 활용하였고 생활 속 이미지의 문자 데이터 세트는 한국어 글자체 이미지 중 Text in the Wild 10만장(표지판·이정표 1.7만장, 상표 3.7만장, 간판 3.0만장, 기타 1.6만장) 이미지를 학습 데이터로 활용하여 각각의 모델 성능을 비교하였다.

그림 2. 학습용 데이터의 예

텍스트로만 구성된 깨끗한 문서를 대상으로 영문 데이터만 학습한 상태에서 영문과 한글 문자영역 탐지율을 표 1과 표 2에서 보여주고 있으며, 한글과 영문 둘 다 학습한 상태에서 영문과 한글 문자 영역 탐지율을 표3과 표4에서 보여주고 있다. 실험결과를 보면 한글과 영문 둘 다 학습한 상태에서 깨끗한 문서의 문자 영역 탐지는 EAST와 YOLO 두 모델이 비슷한 성능을 나타낸다. 반면 표5에서 보여주듯 생활 속 이미지를 대상으로 한 문자열 탐지 결과는 YOLO신경망이 더 나은 결과를 보여주고 있다.

표 1. 영문 데이터만 학습 후 영문 문자열 탐지 결과

모델	Precision	Recall	F1-score
EAST	1	0.9991	0.9996
YOLOv4	0.98	0.99	0.98
YOLOv5	0.792	0.999	0.996

표 2. 영문 데이터만 학습 후 한글 문자열 탐지 결과

모델	Precision	Recall	F1-score
EAST	0.9943	0.9923	0.9933
YOLOv4	0.89	0.97	0.93
YOLOv5	0.823	0.995	0.991

표 3. 영문+한글 데이터 학습 후 영문 문자열 탐지 결과

모델	Precision	Recall	F1-score
EAST	1	0.9978	0.9989
YOLOv4	0.99	1	1
YOLOv5	0.916	1	0.9561

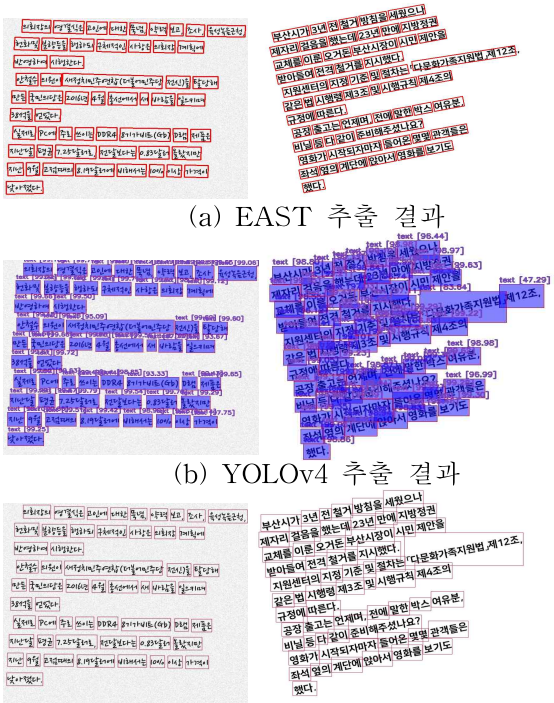
표 4. 영문+한글 데이터 학습 후 한글 문자열 탐지 결과

모델	Precision	Recall	F1-score
EAST	0.9994	1	0.9997
YOLOv4	1	1	1
YOLOv5	0.986	1	0.9929

표 5. 영문+한글 학습 후 생활속 이미지 문자열 탐지 결과

모델	Precision	Recall	F1-score
EAST	0.42	0.35	0.38
YOLOv4	0.64	0.68	0.66
YOLOv5	0.613	0.588	0.6002

EAST의 경우 영문만 학습한 상태에서도 한글문자 영역 탐지율이 높게 나오지만 YOLO의 경우는 한글과 영문 둘 다 학습한 경우 전반적으로 더 나은 문자 탐지 성능을 보이고 있다. 다만 이와 같은 결과는 EAST에 비해 YOLO 신경망이 보다 많은 학습 횟수가 필요하여 나타난 결과일수도 있어서 이에 대해서는 추가적인 연구가 필요하다.



(a) EAST 추출 결과

(b) YOLOv4 추출 결과

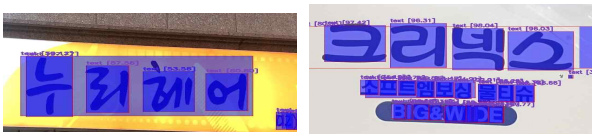
(c) YOLOv5 추출 결과

그림 3. 일반 문서에서 문자열 추출 결과

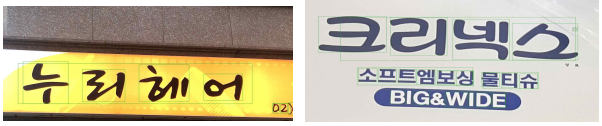
실험결과를 바탕으로 볼 때 두 신경망 모두 문자열 탐지 시 문자열이 가지는 기하학적 특성을 학습하고 이를 바탕으로 문자열을 탐지하는 것으로 유추해 볼 수 있다. 두 신경망이 문자열의 기하하적 패턴을 학습하여 문자열을 탐지하기 때문에 종종 문자열이 없는 이미지에 대해서도 문자 영역을 추출하는 사례가 있는데 그림 5에 이러한 오인식 사례를 보여주고 있다. 그림에서 보듯이 문자열이 전혀 없는 배경을 촬

영한 이미지에 대해서 문자열이 존재하는 것으로 판단하여 단어경계박스를 그리는 것을 알 수 있다.

(a) EAST 추출 결과



(b) YOLOv4 추출 결과



(c) YOLOv5 추출 결과

그림 4. 생활속 이미지 문자열 추출 결과

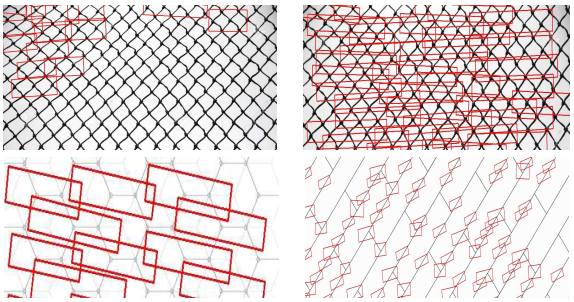


그림 5. 기하학 패턴을 문자열로 오인식한 결과

5. 결론

본 논문에서는 전통적으로 문자열 탐지에서 많이 활용되고 있는 EAST 신경망과 최근 각광을 받고 있는 YOLO 신경망 두 모델을 이용하여 다양한 이미지 속에서 문자열을 탐지하는 실험을 통해 그 성능을 비교해 보았다.

실험 전에는 전통적으로 문자열 탐지 분야에서 많이 사용되는 EAST 신경망 모델이 더 나은 성능을 보일 것으로 예상했으나, 실험결과 일반 텍스트 문서에 대해서는 두 모델이 대등한 성능을 보였으며 생활속 이미지 문자열은 YOLO가 더 나은 성능을 보여주었다. 가볍고 강력한 성능 때문에 다양한 분야에서 활용되는 YOLO는 v4와 v5 최신 모델에서 이처럼 문자 영역 탐지에서도 충분히 강력한 성능을 보여주기에 앞으로 문자인식 기술 분야에도 적극적으로 활용될 것으로 예상된다.

향후 연구에는 추출된 문자영역의 문자들을 인식하는 신경망까지 구성하여 완전한 문자인식 엔진을 구현하고자 한다. 최근 단어나 문장단위로 신경망을 학습하여 문자인식을 수행하는 방법들이 제안되고 또한 우수한 성능을 보여주고 있기에, YOLO 신경망의 문자영역 탐지 기능과 결합한 문자인식 엔진의 성능도 기대해볼 수 있을 것이다.

- 본 연구는 2019~2021년도 중소벤처기업부의 창업성장 기술개발사업 지원에 의한 연구임[S2833775]

참고문헌

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee “Character Region Awareness for Text Detection”, cs.CV 3 Apr 2019

[2] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, Mingxiang Cai “Decoupled Attention Network for Text Recognition”, cs.CV 21 Dec 2019

[3] Zhi Tian, Weilin Huang, Tong He, Pan He, Yu Qiao “Detecting Text in Natural Image with Connectionist Text Proposal Network”, cs.CV 12 Sep 2016

[4] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, Wenyu Liu “TextBoxes: A Fast Text Detector with a Single Deep Neural Network”, cs.CV 21 Nov 2016

[5] Fan Jiang, Zhihui Hao, Xinran Liu “Deep Scene Text Detection with Connected Component Proposals”, cs.CV 17 Aug 2017

[6] ZHOU, Xinyu, et al. East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. p. 5551-5560.

[7] REDMON, Joseph, et al. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 779-788.

[8] Siyang Qin, Roberto Manduchi “Cascaded Segmentation Detection Networks for Word-Level Text Spotting”, cs.CV 3 Apr 2017